**Project Abstract**

**Topic: Extracting significant information form documents using knowledge extraction methods**

**Background:**

Extracting specific information form documents or articles is a manual and slow process that requires a person to first read the document, understand it and then extract the required information from it. This is something that has not changed for some time now.

Extracting information from document soft copies was previously done using parsers. A parser is a program that reads through the whole document line by line similar to how humans do it, and then extract the required information. This automated the process but is slow. Also, if the pattern of the input documents varies, the process becomes slower.

After the introduction of Natural Language Processing (NLP) information could be extracted using a dictionary as reference. This made the process faster, but the model is based on a fixed dictionary that is predefined. This limits the model to adapt to sentences that mean the same but have different words. This leads to decrease in accuracy of the model.

**Proposed system:**

We have planned to develop and train a model that is based on Large Language Model (LLM) to preform word embedding. This defines each word as a vector that is then used as an input to predict its use. We will then develop a classification model on top of this that helps us determine the knowledge contained within the words. This is then used to classify parts of the document into several sections each of which falls into a defined category.

This method improves the accuracy of the result, as it considers multiple uses of all the words in a sentence to categorize it. This will also become faster overtime as it gets more data and training.

**Uses:**

For example,

This model can be used to categorize different parts of a sale deed that can then be used for further processing.